GENERALIZED ENVIRONMENT FOR APPLICATION DEVELOPMENT FOR CAPTURING, EDITING AND CODING STATISTICAL SURVEY'S DATA

Reina Marta Hanono and
Dulce Maria Rocha Barbosa, IBGE, Brasil

ABSTRACT:

The environment enables the generation of client/server or stand alone applications for the entire process of capturing and preparing of survey data. The applications are generated in 'c' source code in order to enable maximum portability across platforms. An application definition facility helps in the creation of the application logic and data base definition. an interactive test facility enables the developers to view both definition and test functions as tightly integrated activities. It allows the incorporation of user's knowledge, improving quality and productivity. It also increases development productivity and performance of the generated applications.

KEYWORDS:

Capturing, Coding, Editing, Imputation, Surving Processing

1. IBGE 'Brazilian Institute of Geography and Statistics'

IBGE is the government agency responsible for all census and statistical surveys in Brazil. It coordinates the National Statistical System, embracing data collection and dissemination for social, industrial, trading, services and agricultural statistics. It records and disseminates data about cartography, geography and natural resources in the country.

Data processing at IBGE covers the demographic census, quinquennial economic and agricultural census, and several yearly and monthly surveys, including the inflation indexes and the economic indicators for all industry sectors and the whole economy. To perform these tasks, millions of questionnaires are collected and processed. IBGE disseminates the results in pre-defined tables (printed or in magnetic form) and performs special queries to meet various public needs.

IBGE has almost 12,000 employees. The planning and development of the surveys and census are centralized at the headquarters and the production is performed by the federal states agencies.

2. INTRODUCTION

The developments in computer hardware and software have given rise to new possibilities for organizing automation in the processing of survey data in statistical bureaus. Following the new trends in automation, we proposed at IBGE a data processing strategy, for the production process for the statistical survey (capturing, coding and editing).Our Goal was to take advantage of new technology, allowing decentralization, portability, independence and integration, in order to bring the subject-matter specialists next to the process of their survey. We focused just on these phases because they were the problematic ones due to the programming and testing consuming resources (time and personal) and specification errors.

With this new strategy the end users will be able to automatically generate, batch or on-line, integrated application programs, enabling execution in a decentralized environment. In that sense, the user must specify, only once, the questionnaire description, the groups of editing rules and the strategy to be followed by the processing, through standard tools which keep automatic documentation in the meta-data enterprise dictionary.

We started with a search for a set of user friendly integrated tools for increasing quality and productivity through the development and the use of the data processing systems for statistical surveys. Finding no appropriate commercial package, we developed a set of tools that can handle any type of questionnaire, and particularly aimed at processing census and surveys.

The set of tools incorporates a proper language called CRIPTAX, which supports the application generation, based on the survey's meta-data definition.

3. SURVEY META-DATA DEFINITION

- ***Survey dictionary***: describes the data physical lay-out (records, variables, range, codes and non-response values), allowing the use of different types of records.

- ***Editing Rules***: describes the rules to be applied in the different steps of processing. The editing rules are written in a standard format using Criptax sentences of the form: IF <condition> THEN<effect>. The **condition** can be a combination of logical expressions and Criptax's functions connected by relational and logical operators. The **effect** is a set of Criptax attribution commands, and the commands **ERR** (to show the record and the associated variables) and **FORGET** (to abandon the record);

- ***Coding Descriptions***: describes the set of codes and texts for the descriptions to be coded (examples: religion, occupation, etc.).

4. TOOLS FEATURES

The package is composed of general tools which perform the functions of survey data processing (capturing, coding and editing), including process control.

The controlling, capturing and coding application generators are menu driven systems. They feature a complete on-line help facility and are independent of any type of programming. They are entirely defined in an interactive mode and interact, at IBGE, with the relational Data Base Management System, TSGBD "Tecnocoop Data Base Management System. They generate applications in a high level language, named OPUS.

The editing program generator runs under IBM mainframe environment and generates 'C' code programs.

In that way the production of the survey's processing can be done in a departmental environment. At IBGE it has been a UNIX one. At present we are moving to a Client/Server architecture supported by UNIX and Windows NT servers, including an IBM 9672 server.

4.1 TOOLS DESCRIPTION

CRIPTAX

Is the editing/correcting tool which enables the user with no data processing knowledge to specify the set of editing rules to be automatically incorporated into the editing application, batch and/or on-line .

It is a pseudo-code style language, composed by procedural blocks delimited by the commands PROC <identification> and END <identification>, where <identification> is a reserved word of the language. These Procs are either pre-defined procedures for which the user must pass the parameters, or are structural blocks where the user specifies the actions to be applied to the data. They allow the user to generate new variables by grouping or recoding variables, previously defined in the dictionary, and the automatic incorporation of the editing rules. The system generates a 'C' code editing application (batch mode) and the editing/correcting application (on-line mode), from the same Criptax editing program.

It provides a user friendly tool, fully integrated with the Institutional Meta-Data Base for the specification of the questionnaire structure and editing rules. It enables cross record processing of the survey data, which can reside in different relational Data Base management Systems.

SISENT

Is the data entry tool which allows the generation of :

- Survey Database Design based on the survey dictionary;
- Record screens definition for entering the questionnaire's fields.

  Each screen must contain variables of a particular record type. Automatic checks for the data will be performed, during data capture, according to the dictionary specifications. Optionally the user can define consistency functions in Criptax, to be performed during data capture. The system allows screen testing during the definition phase.
- Access routine to the survey Data Base, allowing record insertion, deletion and updating;
- Manager program for the capturing application.

SISCOD

Is the Coding Application Generator, an interactive specification facility which enables the generation of:

- Coding Data Base Design, based on the coding descriptions and parsing options, specified by the user;
- Coding testing;
- Coding Data Base update program;
- Survey Data Base Design , which uses the survey dictionary (optional);
- Coding environment which is based on the Survey and Coding Data Base designs;
- Automated Coding Application (batch mode);
- Computer-assisted Coding Application (on-line mode);
- Consistency functions based on the editing rules defined in Criptax.

SISCON

Is the Controlling Application Generator which allows the generation of :

- Controlling Data Base design;

•Controlling application, which is composed of the access and update routines to the controlling Data Base, which can be used in each phase of the survey processing;
• Controlling reports' programs;
• Manager program to command controlling reports.

5. METHODOLOGICAL PROCESSING CHANGES

The set of tools had introduced a change in the way of working. They allow for participative development, in which the users with no data processing knowledge, subject matter specialists, are directly responsible for the specification of the editing rules, as far as for the parsing options for the coding application. In this way the users' knowledge about editing and coding is incorporated during the development phase.

These facts necessarily changed the working methodology, specially for the thematic specialists, who are now responsible for the application correctness leaving to the informatic specialists the responsibility for application performance and feasibility.

Specific training was required for the development team at the different phases of the process. It consisted in training for the specification of the Data Dictionary and Editing Rules using Criptax language. The informatic specialists were trained in the specification of Criptax programs and routines for the editing rules that couldn't be defined directly by the users. Very efficient and effective skills transfer modules were developed and they always took less than fifteen hours.

6. ADVANTAGES

The most important advantage of the set of tools is that they perform the major task in survey data processing allowing the users with little or no computer experience to define the requirements for the application.

In summary the advantages are:
• easy to use and learn;
• menu driven operation for controlling, capturing and coding generators;
• tools integration, that is, what is obtained by one phase can immediately be used for the next without having to redefine the dictionary and redesign the Data Base;
• prototype development methodology, incorporating user's knowledge;
• portable application generation;
• minimum need of programming effort, only for the editing application;
• reduction of application development maintenance and cost;
• improve of quality;
• improve performance.

7. USAGE AT IBGE

CRIPTAX has been used for the generation of the editing application for the Demographic Census 1991, the National Household Surveys ('PNAD') 92, 93 and 95 in a UNIX environment and for the Agricultural and Economical Census 95 in a Client/Server architecture.

SISENT has been used for the generation of Data Capture Application for several monthly surveys (Industrial, Milk Production, Building Activities, etc.) in a UNIX environment.

SISCOD has been used for Coding in the Demographic Census 1991, the National Household Survey 92, 93 and 95 and the Familiar Budget Survey (‘POF ‘) 1995.

In all cases, we realized a significant reduction of time for application development, reduced specification errors, and increase user satisfaction.

8. FUTURE DEVELOPMENTS

IBGE is moving to a Client/Server architecture with Oracle Data Base Management System and Windows. We plan to transform the existing tools for controlling, capturing and coding,currently running under UNIX, to this platform taking advantage of the graphics facilities offered by this new environment, providing a more friendly user interface.

The generated editing application are easily transferred to this new environment since they are actually generated in ‘C’ code and the interface routine with the survey’s data base is supplied by the user. Nevertheless we are planning the development of a more friendly user’s interface for the editing program specification. We are also analyzing the cost/benefit of adapting the actual generator running on an IBM mainframe, to a new one running under Windows.

GLOSSARY:

coding - the act of attributing a numerical code associated to a text description;

consistency functions - a 'C' routine generated by Criptax system based on a set of editing rules. This routine receives a variable as a parameter and performs all the editing rules related to this variable.

correction - the action of changing the value of a variable that was considered invalid or inconsistent according to the editing rules. It can be automatic or manual;

editing rule - a condition that detects an error for a record or a questionnaire;

editing - the act of applying a set of editing rules and corrections so as to produce a clean file;

REFERENCES:

1 BARBOSA, D.M.R. and HANONO, R.M., Estudo das ferramentas para apuração de dados. Revista Brasileira de Estatística, Rio de Janeiro, 49(191) Jan/Jun 1988, pag. 85-100.

2 CRIPTAX - Users Manual - IBGE *

2 HANONO, R.M. and BARBOSA, D.M.R., A Tool for the Automatic Generation of Data Editing and Imputation Application for Surveys Processing. Survey and Statistical Computing (SGCSA) - North Holland - pag. 449-456.

10 OPUS - Programming Manual - Tecnocoop Sistemas Brasil

3 SILVA, A.C.M, HANONO, R.M. and BARBOSA, D.M.R., A Tool for the Automatic Generation of Computer-assisted Coding Application. Submitted and accepted by SGCSA - Computerising Survey Support Systems - 1993.
4 SISCON - Controlling Application Generator - User's Manual - IBGE *
4 SISCOD - Coding Application Generator - User's Manual - IBGE *
5 SISENT - Capturing Application Generator - User's Manual - IBGE *
9 TSGBD - Data Base Management System - Tecnocoop Sistemas Brasil *

* Papers from IBGE are available in portuguese at:
IBGE - Diretoria de Informatica - Biblioteca
Rua Visconde de Niteroi, 1246 CEP: 20943-001
Rio de Janeiro - RJ - BRASIL
FAX (5521)(2484123) (2849598)